LANGUAGE TECH NEWS

ata A Publication of the Language Technology Division of the American Translators Association.

From the Assistant Administrator:

Welcome to another great issue of the Language Technology Division Newsletter. Once again we have a variety of articles for your reading pleasure, not just on translation tools, but on broader subjects as well.

Due to the way ATA divisions work, most of a division administrator's activities are focused on the annual conference. Some language divisions have a mid-year conference, but the tools seminar held this year in San Francisco in March was organized by the ATA Professional Development Committee, not by the LTD. Indeed, many tasks that would perhaps normally fall to the LTD, such as evaluation of proposed talks for the annual conference, are actually performed by the ATA Translation and Computers Committee, which predates the LTD. (I am also on the Committee, for full disclosure!)

One thing the LTD does is to find "distinguished speakers" for the annual conference who then have to be approved by the conference organizer in order to be reimbursed for travel and hotel expenses. I have been working hard and have found two outstanding speakers, Dr. Lisa Sattler, a specialist in physical therapy, who will talk about office ergonomics, and Prof. Klaus Dirk Schmitz of the Cologne University of Applied Sciences, who will talk about terminology. See page 2 for details about their talks.

Note that the distinguished speakers have to be individuals who do not normally attend ATA meetings, and this usually means non-ATA members and foreign linguists. Last year's speaker was a monolingual computer repair technician, Carey Holzman. He gave many tips on how to deal with the Windows OS, his specialty.

VOL. 3, NO. 1 / JULY 2009

IN THIS ISSUE:

Conference Speakers 2
Call for Nominations 3
Controlled Language: Does my Company Need It? 4
Found CAT 7
Trados Tip 9
A Survey of Corpus Tools for Translators 13

Outside of the annual conference, our only opportunities to talk to one another and share our knowledge are through the newsletter, the mailing list and the blog on the LTD site. Please visit the LTD site if you have not been there recently! And if you are interested in contributing to the blog, or just want to forward information for me or Michael Wahlster <mwahlster@gw-language.com> to post on the blog, please send one of us an email.

This year is an election year for the LTD, so please see the information on nominations on page 3 of this newsletter and please send in a nomination to the nominating committee! We need you to participate.

Naomi Sutcliffe de Moraes
Assistant Administrator

Distinguished Speakers: 2009 Conference

Dr. Lisa Sattler

The Importance of Ergonomics for Translators: How to avoid repetitive strain injuries

Friday, October 30, 2:00-3:30pm

The injuries people get from sitting long hours at a computer are usually called repetitive strain injuries. A person may receive one of many diagnoses, including tendinitis and carpal tunnel syndrome. These overuse injuries are prevented and corrected in the same way. Good posture and ergonomics work together to aid prevention and correction. This 90 minute lecture will discuss the signs and symptoms of more common injuries to help you recognize them before they become severe. It will include information about what you can do to heal or prevent injuries, including detailed ergonomic recommendations, posture training and stretches.

Prof. Klaus-Dirk Schmitz

Cologne University of Applied Sciences

Terminology management for localization of software user interfaces

Thursday, October 29, 2:00-3:30pm

The localization of software products has to deal with different types of text, such as installation manuals, on-line help files, packaging and marketing material, websites, and the software user interface. While the first text types are more or less typical technical texts, the user interface—with menus, dialog boxes, tool tips and error messages–requires a dedicated approach for terminology management. This session will demonstrate terminological phenomena typical of software user interfaces, discuss the value of traditional terminology management for this kind of technical text, and develop a proposal for an adequate terminological data model.

Learn from terminology standards: How can freelance translators and small language service providers set up a detailed, practical terminology management solution.

Saturday, October 31, 2:00-3:30pm

International terminology standards such as ISO 16642 (TMF), ISO 12620 (DatCats) and ISO 30042 (TBX), as well as established best practices, provide a set of principles and methods for setting up a terminology management system. The language and translation departments of huge industrial companies and public organisations are not the only ones who can benefit from these guidelines. Small language service providers and freelance translators should also make use of this professional know-how. This session will give a short theoretical background on terminology management, explain basic design principles and typical data categories for termbases, and show how terminology management systems can be used to support translators.

Language Tech News

Vol. 3, No. 1 July 2009

Copyright©2009

American Translators Association
225 Reinekers Lane, Suite 590
Alexandria, VA 22314
Telephone (703) 683-6100
Fax (703) 683-6122
ata@atanet.org
www.atanet.org

Editor
Roomy Naqvy
roomynaqvy@gmail.com

Editorial Committee
Naomi J. Sutcliffe de Moraes
Barbara Guggemos

Proofreader:
Naomi J. Sutcliffe de Moraes

Contributors to this Issue:
Tuomas Kostiainen
Naomi J. Sutcliffe de Moraes
Uwe Muegge
Thelma L. Sabim

Layout:
Cindy Gresham
illustrator@cgresham.com

LTD is the Language Technology Division of the American Translators Association

LTD Administrator:
Dierk Seeburg
info@icotext.com

LTD Assistant Administrator:
Naomi J. Sutcliffe de Moraes
www.justrightcommunications.com

Call for Nominations

The Language Technology Division is pleased to call for nominations from the LTD membership for the following positions:

Administrator (2-yr term)

Assistant Administrator (2-yr term)

Election of these officers is held every two years in accordance with the LTD bylaws. The results of the election will be announced at the LTD Annual Meeting, which will be held during ATA's 50th Annual Conference in New York City, October 28-31, 2009.

LTD Officer Duties

Officers must be members of the Language Technology Division as well as voting members of ATA. You will find a summary of duties for both the administrator and assistant administrator positions online at: http://www.americantranslators.org/divisions/Officer_Duties.pdf

Serving in a division leadership role provides enormous opportunity, both professionally and personally. Division officers frequently find themselves becoming more successful in their own careers as they develop additional skills, make useful business connections, and share ideas with other division members.

How to Nominate a Candidate

Your assistance in helping the LTD Nominating Committee identify interested, capable colleagues is crucial to the election process and the division. Qualified candidates must be voting (active or corresponding) members of ATA and members of the Language Technology Division. Any division member may make a nomination, and self-nominations are also welcome.

If you plan to put a name forward for a nomination, it would be helpful if you could contact the potential nominee first and tell them of your intention. Let them know that a nomination does not guarantee a formal invitation to run for office. Remember that LTD officers serve on a volunteer basis; please do not nominate colleagues who express serious concerns about service or who have conflicting priorities.

To nominate a candidate for a LTD office, you may contact any member of the Nominating Committee listed below or download the Nomination Form from http://www.ata-divisions.org/LTD/wpcontent/uploads/nom_form_ltd31mar09.doc. The nomination form may be mailed or faxed to ATA Headquarters.

LTD Nominating Committee

A Nominating Committee has been appointed to actively seek nominations for candidates. Members of the 2009 ID Nominating Committee are:

Betty Welker
(Bettywelker@cs.com)
Jost Zetzsche
(jzetzsche@internatinalwriters.com)

Election Schedule

July 24	Slate of candidates published
Sept.7	Deadline for receipt of petition to add candidates to slate
Sept.18	Ballots mailed if more than one candidate is running for any office
Oct. 23	Deadline for receipt of ballots

We hope you will take this opportunity to consider stepping forward as a volunteer during the coming year – if not as a candidate for office, then perhaps as a contributor to our division newsletter or by giving a talk at the annual conference. There are many ways to be involved, and volunteering is a wonderful way not only to share your experience but also to expand your network of contacts. As always, your support of the Language Technology Division and ATA is appreciated. ▢

Thank you,
2009 LTD Nominating Committee

Controlled Language: Does My Company Need It?

By Uwe Muegge

Controlled languages use basic writing rules to simplify sentence structure. Here is how they work and how your company can benefit from introducing a controlled language.

Editor's Note: This article has been reprinted with permission of tcworld magazine, www.tcworld.info and it can be accessed at www.tcworld.info/file/tcworld_2009_02.pdf

What is a controlled language?

A controlled language is a subset of a natural language, as opposed to an artificial or constructed language. Natural languages such as English or German are languages that are used by humans for general communication. A controlled language differs from the general language in two significant ways:

1. The grammar rules of a controlled language are typically more restrictive than those of the general language;
2. The vocabulary of a controlled language typically contains only a fraction of the number of words that are permissible in the general language.

As a result, authors who use a controlled language have fewer choices available when writing a text. For example, the sentence "Check the spelling of a paper before publishing it" is a perfectly acceptable sentence in general English. Using CLOUT™, a controlled language rule set developed by the author of this article, the sample sentence would have to be rewritten as "You must check the spelling of your document before you publish that document" to comply with rules regarding vocabulary, active voice, use of articles, and avoidance of pronouns.

Why do we need controlled languages?

Facilitating language learning

Probably the first controlled language, *Basic English* was created by C.K. Ogden in 1930[i]. The developer had the explicit goal of dramatically reducing the 5+ years it takes to master Standard English. Based on a vocabulary that contains 850 essential words (the Oxford English Dictionary, on the other hand, defines more than 600,000 words), *Basic English* is designed to be acquired in just a few weeks.

Eliminating translation

One of the most widely used controlled languages today is ASD-STE100 *Simplified Technical English*[iii], also known as *Simplified English*. *Simplified English* was originally developed by the European Association of Aerospace Manufacturers (AECMA) in the 1980s. The main purpose of *Simplified English* was to create a variant of Standard English that aircraft engineers with only a limited command of English could understand, thereby eliminating the need to translate maintenance manuals into foreign languages.

Streamlining translation

Within the localization industry, many people familiar with the controlled language concept associate controlled language with automating the translation process. In fact, it typically comes as a surprise that controlled languages can and have been used for purposes other than making the translation process more efficient. By restricting both vocabulary and style, using a controlled language typically improves match rates in translation memory environments and trans-

Uwe Muegge is the Director of MedL10N, the life science division of CSOFT. He is currently a member of TC37 at the International Organization for Standardization (ISO) and teaches graduate courses in Terminology Management and Computer-Assisted Translation at the Monterey Institute of International Studies. Uwe can be contacted at uwe.muegge@medl10n.com Visit his website at www.medl10n.com

Examples of organizations that have created a controlled language:

Alcatel: Controlled English Grammar (COGRAM)

Avaya: Avaya Controlled English (ACE)

Caterpillar: Caterpillar Technical English (CTE), Caterpillar Fundamental English (CFE)

Dassault Aerospace: Français Rationalisé

Ericsson: Ericsson English

General Motors (GM): Controlled Automotive Service Language (CASL)

IBM: Easy English

Kodak: International Service Language

Nortel: Nortel Standard English (NSE)

Océ: Controlled English.

Siemens: Siemens DokumentationsDeutsch

Scania: Scania Swedish.

Sun Microsystems: Sun Controlled English

Xerox: Xerox Multilingual Customized English

lation quality in (rule-based) machine translation environments.

Enhancing comprehensibility

Helping authors avoid both semantic and syntactic ambiguity has been recognized as a goal worth pursuing in and by itself, especially in the domain of technical communication. Some organizations are deploying a controlled language for the sole purpose of improving the user experience of a product or service in the domestic market.

Common features

One characteristic that most controlled languages share is the fact that very little information about their rule sets and vocabularies is freely available. This is not really surprising when you consider the fact that a controlled language holds the promise of giving the organization that uses it a distinct advantage over its competition.

The other feature many controlled languages have in common is their dissimilarity. Nortel Standard English, for instance, has only a little over a dozen rules, while Caterpillar Technical English consists of more than ten times as many. A recent comparative analysis of eight controlled English languages found that the number of shared features was exactly one, i.e. a preference for short sentences.[iv]

Why should my organization use a controlled language?

Improved usability

Documents that are more readable and more comprehensible improve the usability of a product or service and reduce the number of support calls.

Objective metrics and author support

Tools-driven controlled language environments enable the automation of many editing tasks and provide objective quality metrics for the authoring process. Controlled language environments also provide authors with powerful tools that give them objective and structured support in a typically rather subjective and unstructured environment.

Lower translation costs

As controlled language texts are more uniform and standardized than uncontrolled ones, controlled language source documents typically have higher match rates when processed in a translation memory system than uncontrolled source documents. Higher match rates mean lower translation costs and higher translation speed.

Some controlled languages have been specifically designed with machine translation in mind, e.g. *Caterpillar Technical English* or this author's *Controlled Language Optimized for Uniform Translation CLOUT*. Using a controlled language customized for a specific machine translation system will significantly improve the quality of machine-generated translation proposals and dramatically reduce the time and cost associated with human translators editing those proposals.

Impact on translation?

Status quo

One of the biggest challenges facing organizations that wish to reduce the cost and time involved in the translation of their materials is the fact that even in environments that combine content management systems with translation memory technology, the percentage of untranslated segments per new project can remain fairly high. While it is certainly possible to manage content on the sentence/segment level, the current best practice seems to be to chunk at the topic level. Chunking at the topic level means that reuse occurs at a fairly high level of granularity. In other words: There is too much variability within these topics!

Controlled authoring for translation memory systems

Writing in a controlled language reduces variability, especially if the controlled language not only covers grammar, style, and vocabulary, but also function. In a functional

approach to controlled language authoring, there are specific rules for text functions such as instructions, results, or a warning message. Here are two simple examples for functional controlled language rules:

Text function: Instructions
Pattern: Verb (infinitive) + article + object + punctuation mark.
Example: Click the button.

Text function: Results
Pattern: Article + object + verb (present tense) + punctuation mark.
Example: The window "Expense Report" appears.

Implementing functional controlled language rules will enable authors to write text where sentences with the same function have a very high degree of similarity. This not only makes sentence modules reusable within and across topics in a content management system, it also dramatically improves match rates in a translation memory system.

Controlled authoring for rules-based machine translation systems

Unlike in a traditional translation memory environment, where uniformity is the decisive factor in improving efficiency, the big factor for making machine translation systems more productive is reducing ambiguity in the source text. The problem that rules-based machine translation systems like Systran struggle with is the fact that in uncontrolled source texts, the (grammatical) relationship between the words in a sentence is not always clear. To enable rules-based machine translation systems to generate better translations, the controlled language needs to have rules like the following that helps the machine translation system successfully identify the part of speech of each word in a sentence:

Write sentences that have articles before nouns, where possible.

Do not write: Click button to launch program.
Write: Click the button to launch the program.

Write sentences that repeat the noun instead of writing a pronoun.

Do not write: The button expands into a window when you click it.
Write: The button expands into a window when you click the button.

With rules in place that mitigate the weaknesses of rules-based machine translation systems, the quality of the output produced by these machine translation systems is bound to improve dramatically.

Do I have to develop my own controlled language?

Not at all! Today, many organizations that wish to reap the benefits of controlled-language authoring opt for a software-driven solution that comes with a built-in set of grammar and style rules. Systems like acrolinx *IQ Suite*, IAI *CLAT*, or Tedopres *HyperSTE* have enabled literally thousands of organizations to improve the quality and productivity of their authoring and translation processes. In a software-driven authoring environment, organizations do not have to maintain the staff of highly trained linguistic experts needed to develop and deploy a proprietary controlled language. Instead, the organization simply selects the rules that are most suitable for a given content type from a set of preexisting writing rules. Typically, these checking tools support the definition of multiple sets of rules for multiple types of content (e.g. stricter rules for user documentation than for knowledge-base articles).

Terminology management support

From a technology standpoint, it is relatively easy to implement the rules part of a controlled language, the terminology part is typically more labor intensive. It is certainly true that many controlled language software solutions include a module for collecting terminology. However, the task of creating a corporate dictionary, which is what this job amounts to, might be daunting. Not only will all synonyms, among the possibly thousands of

terms in use at the organization, have to be identified, but these synonyms will also have to be categorized into preferred and deprecated (do not use) terms. While creating a corporate dictionary may be a challenge, once it is available, that dictionary may also be the feature most valued by the users of the controlled language system.

Example of a controlled language

To see an implementation of a simple controlled language designed for machine translation, visit the author's website at www.muegge.cc. The entire site was written in *CLOUT*, the Controlled Language Optimized for Machine Translation. On the home page, click on any of the language combinations into English, i.e. German > English or French > English and watch how *Systran*'s free machine translation system turns a complete website into a fully navigable, highly comprehensible virtual English site in real time. Click on the link *Controlled Language/Rules for Machine* to see ten sample *CLOUT* writing rules that have a high impact on the comprehensibility and (machine) translatability of instructional text in English. ▢

Notes:

i Ogden, Charles Kay. 1930. Basic English: A General Introduction with Rules and Grammar. London : Treber, 1930.

ii Basic English Institute. 1996. Ogden's Basic English Word List. Ogden's Basic English. [Online] 1996. [Cited: February 3, 2009.] http://ogden.basic-english.org/words.html.

iii AeroSpace and Defence Industries Association of Europe. 2005. ASD-STE100 - Simplified Technical English - International Specification for the Preparation of Maintenance Documentation in a Controlled Language. ESSAS Electronic Supporting System for ASD Standardization. [Online] 2005. [Cited: February 3, 2009.] http://www.asd-stan.org/sales/asdocs.asp.

iv O'Brien, Sharon. 2003. Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets. Machine Translation Archive. [Online] 2003. [Cited: February 3, 2009.] http://www.mt-archive.info/CLT-2003-Obrien.pdf.

Found CAT

By Thelma L. Sabim

There are no fewer than twenty Computer-Assisted Translation or CAT programs out there. So how is one to choose? As I see it, the three main factors are cost, ease-of-use and compatibility.

Given the proliferation of new file types, a program ought to be flexible enough to accommodate at least most of them, without constantly bleeding your pocketbook for upgrades. My favorite is Omega T. It is open-source, available free of charge at: http://sourceforge.net/projects/omegat.

Since its official release in 2002, Omega T has attracted many developers and contributors, no small fraction of whom are translators. These are people living all over the world, and running computers on different platforms. Their suggestions are posted on a Yahoo users list (http://tech.groups.yahoo.com/group/OmegaT). There, you can look at a given problem through the eyes of a Mac user in Japan, a Linux user in Germany and a Windows PC user in Russia. It is a trader's bazaar of knowledge, experience and suggestions from all over the world, staffed entirely by volunteers.

The first step in OmegaT is to create a project. Set up your source and target languages and accept the default folders. Later, when you

Thelma L. Sabim, a native of Brazil, has been working as a full-time freelance translator in Austin, Texas and Curitiba, Paraná since 1989. She is a certified translator in the USA and Brazil and a volunteer localizer of this open-source CAT tool. She can be contacted at: tlsabim@portuguesetranslator.net

get more familiar with the program, you can fiddle with these settings. (See Fig. 1)

The source files need to be placed in the source folder. Microsoft Office files must be converted into OpenOffice formats first. (See Fig. 2)

Any translation memory—in tmx format—will go into the TM folder. The number of TMs is limited by the power of your PC. Technically you can use as many TMs as you (or your PC) want.

The option of including **multiple** *file types* and *pre-existing memories* within a project helps ensure consistency: I can see how I translated a phrase on a slide, and with a single click, can see how the same sentence came out in the manual.

Fig. 3 shows the segment # 0055 "Fuzzy matching" opened for translation and the matches available in the TMs. I can type Ctrl+1 to highlight Option 1 or Ctrl+2 to highlight Option 2. Then Ctrl+I inserts the selected match in the opened segment.

OmegaT works with pre-existing glossaries, too. The file needs to be in three columns and in the tab-delimited format. I do not have experience working with glossaries in OmegaT, but the User's Manual has detailed information about this feature.

When the translation is finished, the next step is to save the project and then create the target file(s). It is a good idea to save the project and create the target file also during the translation process (Fig. 4, page 20). To view translated documents during translation, just click on the files in the target folder. (Fig. 2)

The last step is to convert the two files back into Microsoft format, if necessary.

OmegaT's compatibility with different *platforms* is a selling point to me, because I'm planning to move to Linux. It means I won't have to spend a lot of time learning new CAT programs that only run on one operating system.

Another compatibility plus is that OmegaT doesn't "pick fights" with memory-hungry speech recognition programs like Dragon and ViaVoice.

cont. page 20

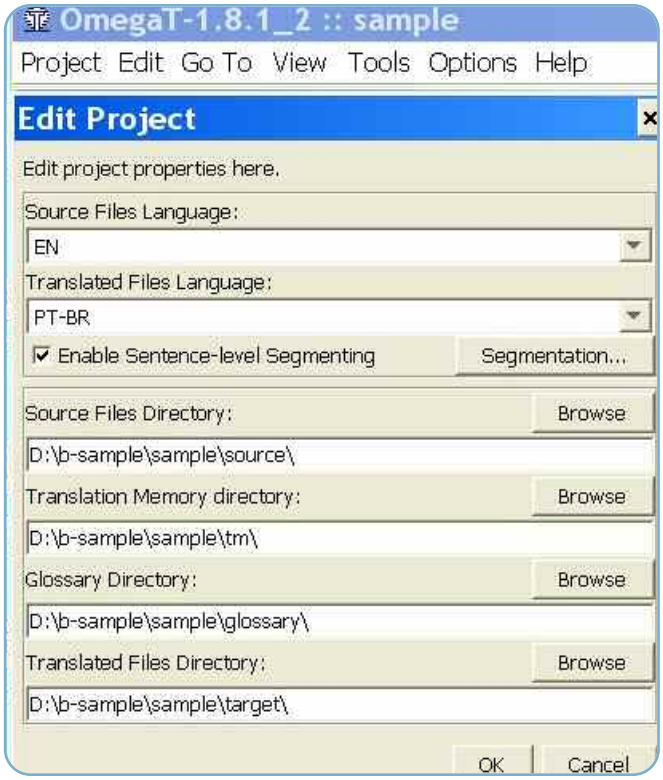


Figure 1, above

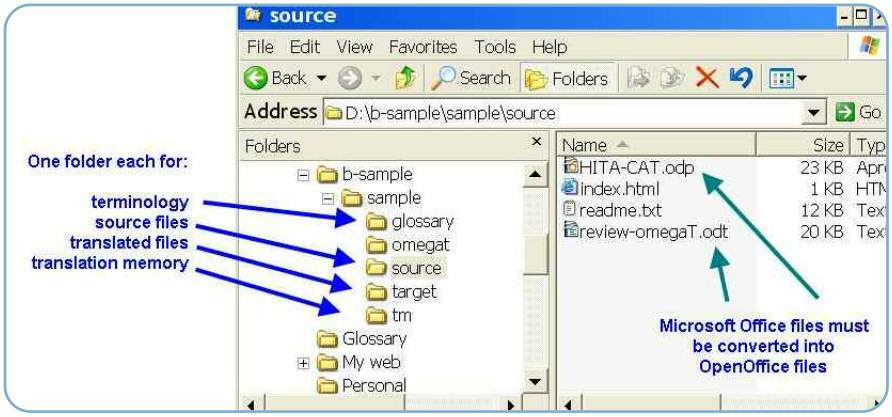
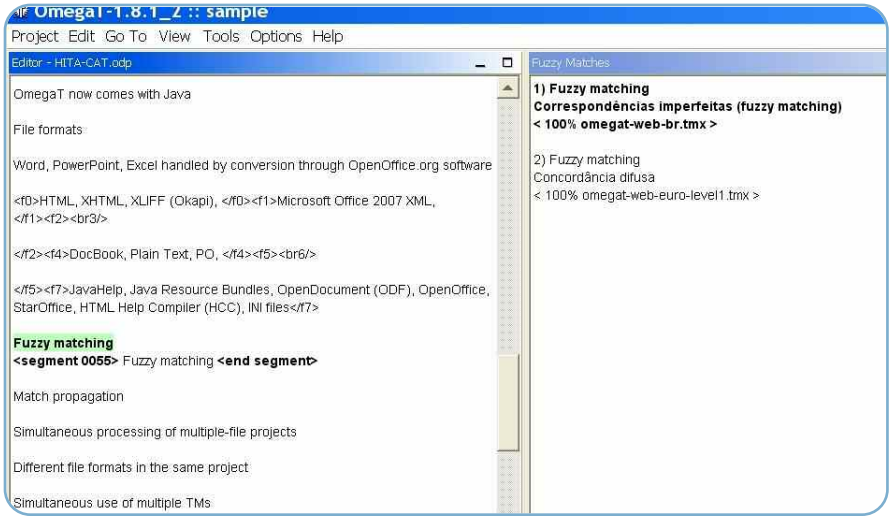


Figure 2, above

Figure 3, below


OmegaT-1.8.1_2 :: sample
Project Edit Go To View Tools Options Help
Editor - HITA-CAT.odp
Fuzzy Matches
OmegaT now comes with Java
File formats
Word, PowerPoint, Excel handled by conversion through OpenOffice.org software
<f0>HTML, XHTML, XLIFF (Okapi), </f0><f1>Microsoft Office 2007 XML, </f1><f2><br3/>
</f2><f4>DocBook, Plain Text, PO, </f4><f5><br6/>
</f5><f7>JavaHelp, Java Resource Bundles, OpenDocument (ODF), OpenOffice, StarOffice, HTML Help Compiler (HCC), INI files</f7>
Fuzzy matching
<segment 0055> Fuzzy matching <end segment>
Match propagation
Simultaneous processing of multiple-file projects
Different file formats in the same project
Simultaneous use of multiple TMs
1) Fuzzy matching
Correspondências imperfeitas (fuzzy matching)
< 100% omegat-web-br.tmx >
2) Fuzzy matching
Concordância difusa
< 100% omegat-web-euro-level1.tmx >

Trados Tip

by Tuomas Kostiainen

Using MultiTerm With Trados

Part 2: Where Do I Get MultiTerm Glossaries?

In my previous article (January 2009), I explained how MultiTerm is used with Trados and how the Term Recognition feature works in Trados. That's really good and important, but if you don't know how to create or convert MultiTerm glossaries, it's all quite useless. I know that you all have been extremely anxious to get your hands on this follow-up part so that you can put that great feature into practice.

So, how do you get those MultiTerm termbases? You basically have the following three ways of getting them:

1. By creating a new termbase from scratch
2. By converting an existing glossary from some other file format
3. By loading a MultiTerm termbase in MDB format into your own termbase library

In this article, I will concentrate on the Methods 2 and 3 because they are the most common, and even if you want to create a glossary from scratch, this is often easier to do using Method 2. If you want to create a new termbase and initially have only a small number of terms to add, you can do it directly in MultiTerm. However, if you are planning to enter numerous terms into a new glossary, you might want to create the glossary first in Excel and then import it into MultiTerm. The reason is that entering a large number of terms is faster in Excel than in MultiTerm.

Converting Terminology Data to MultiTerm (XML) Format

SDL MultiTerm comes with a separate application called SDL MultiTerm Convert which allows you to convert your non-MultiTerm glossaries into MultiTerm termbases. You can convert the following file formats:

- MultiTerm 5 format
- Olif XML format
- SDL Termbase Desktop format
- SDL Termbase Online format
- Spreadsheet or database exchange format
- Microsoft Excel format

The conversion is a three-step process. First, you create an ***.xml** and ***.xdt** file from the source file using MultiTerm Convert. The XML file is the termbase **data** file that includes the actual terminology data, while the XDT file is the termbase **definition** file that contains the structure of the termbase. The second step is to create a new termbase in MultiTerm based on the XDT file, and the third step is to import the data in the XML file into the new termbase. This might sound a wee bit complicated, and I must admit that it is a more complicated process than it should be; however, if you know what to do, it works quite quickly and smoothly.

Regardless of the format of the source file, the last two steps of the process are always the same. Only the first step (conversion to XML and XDT files), which is done in MultiTerm Convert, varies depending on the source file type. Here, I will explain how the conversion process is done with Excel files, because this is the most common file format for conversion (and many other formats, such as Word tables and csv files, can easily be converted to Excel format). You can find additional information regarding the other file formats in the MultiTerm User Guide or Online Help.

Converting Excel Data

Step 1: Convert data

1. Prepare your glossary file in Excel so that each column has a header on the first row and the source and target term fields include only terms/phrases and no explanations, synonyms, alternative endings or other information. All the other information can be placed in separate columns, which should then be labeled accordingly. All the data has to be on the first worksheet of the file and there should be no empty columns between the columns that contain the data. For example, see Figure 1.
2. Open MultiTerm Convert (**Start > All Programs > SDL International > SDL MultiTerm 2007 > SDL MultiTerm 2007 Convert**).

	A	B	C	D	E
1	English	Finnish	Notes	Sample species	Esimerkkilaji
2	Butterfly	perhonen	2 pairs of wings	Monarch	monarkki
3	Fly	kärpänen	1 pair of wings	Housefly	huonekärpänen
4	Mite	punkki	not an insect	Dust mite	pölypunkki
5	Spider	hämähäkki	not an insect, 8 legs	Black widow	mustaleski
6	Beetle	kovakuoriainen	2 pairs of wings	Tiger beetle	hietakiitäjäinen
7	Ant	muurahainen	about 10,000 species in the world	Common red ant	kekomuurahainen
8	Aphid	kirva	pear-shaped body	Black bean aphid	juurikaskirva
9	Bumble bee	kimalainen	large hairy social bee	Common carder bee	peltokimalainen
10					

Figure 1. A properly structured sample glossary in Excel format. This glossary includes 8 terms (rows 2 - 9) and 5 fields. The field names are on the first row.

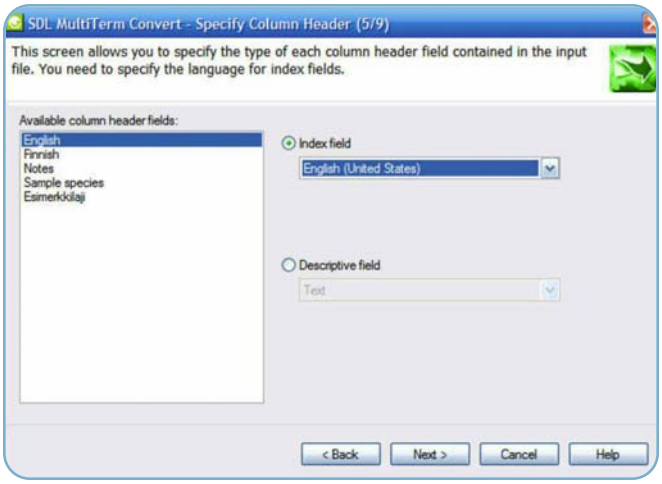


Figure 2. Defining Index fields or Descriptive fields. Each of the 5 fields has to be defined as an Index field or a Descriptive field. Note that MultiTerm does not automatically define a field as an index (= language) field even if the field name (such as "English") is clearly a name of a language.

3. Specify conversion session options (New/Save/Existing). Select **New conversion session** if you do not have a previous session that you would like to reuse. You can save your new conversion session by selecting **Save conversion session** and giving a name and folder for the session file. Note that this is only the conversion session file and not the actual termbase. In most cases, there is no need to save the session. If you want to reuse an existing session instead of creating a new one, select **Load existing conversion session**. After you have selected the options, click **Next**.
4. Specify the file type of the source file. Since we are converting an Excel file, select **Microsoft Excel format**. Click **Next**.
5. Specify the input file in the **Input file** box by clicking **Browse** and then locating the Excel file that you want to convert. The other three file names will be filled out automatically and the files will be placed into the same folder where your input file is. Click **Next**.
6. Specify which ones of the **Available column header fields** are index fields (= languages) and which ones are descriptive fields. Do this by selecting one of the listed header fields and then selecting either the **Index field** or **Descriptive field** radio button. Let's say you used "English" as the column header for your English term column. Select "English" in the list of column headers, then select the **Index field** radio button and select **English** from the pull-down menu. Do this for all the other language fields in the glossary (See Figure 2).
7. Next specify the descriptive fields. All descriptive fields are text fields by default. If your descriptive fields are not text fields, you can change their field type by first selecting the field in question in the **Available column header fields** list and then selecting the appropriate field type from the pull-down menu under the **Descriptive field** button. The available field types are **Text**, **Number**, **Boolean**, **Date**, **Picklist**, and **Multimedia file**. When you have specified all the fields click **Next**.
8. Create the entry structure by adding the descriptive fields to their "correct" locations within the structure. This is really your own decision and depends on your glossary structure. If you are unsure where the fields should go, just place them somewhere in the structure. You will see later how logical (or illogical) the locations were and can change them if needed. The location of a field within the glossary structure does not affect how MultiTerm works with Trados Workbench. To add a field to a location, select one of the descriptive fields under **Available descriptive fields** and then select the location in the **Entry structure** where the field should go and click **Add**. Do this to all the descriptive fields. You can also remove a field from the structure by selecting it and clicking **Remove**. Note that a field can be inserted into more than one location in the entry structure. When you are satisfied with the structure, click **Next** (See Figure 3).
9. The **Conversion Summary** window gives you a summary of the files and their locations. Check that **Convert immediately** is selected and click **Next**.
10. Check how many "entries were successfully converted" in the **Converting** window. It should match the number of entries in your Excel file. Click **Next** and then **Finish**.

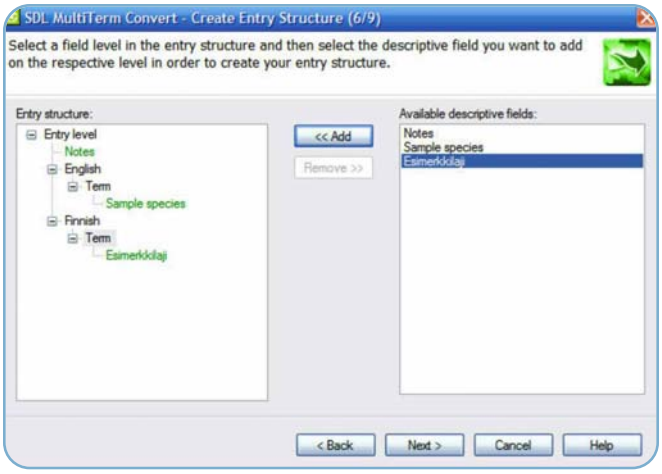


Figure 3. Create a termbase entry structure by adding descriptive fields to their "correct" locations within the structure. Here "Notes" field has been placed on the top, "Sample species" under the English term, and "Esimerkkilaji" (Finnish sample species) under the Finnish term. For the resulting entry structure, see Figure 4.

The converted data (XML file) is now ready to be imported into a MultiTerm termbase. However, first you need to create a termbase into which to import the converted data. *See Step 2: Create a new termbase.*

Step 2: Create a new termbase

1. Open MultiTerm (**Start > All Programs > SDL International > SDL MultiTerm 2007 > SDL MultiTerm 2007**).
2. Select **Termbase > Create Termbase**.
3. Choose termbase location. You might want to create one specific folder for all your MultiTerm glossaries.
4. The **Termbase Wizard** window comes up. Run the Wizard by selecting **Next**.
5. In the Termbase Definition window, select **Load an existing termbase definition file** option and locate the XDT file (= termbase definition file) that was created during the conversion session in Step 1.
6. Click **Next** and enter a name for the termbase in the **Termbase Name** window under **Name**. You can also add optional description and copyright information here. If you click **Add More** you can even add a splash screen and icon for the termbase. Click **Next**.
7. On the **Index Fields** page, verify the index fields information. This should be correct because it is based on the definition you created during the conversion session. Click **Next**.
8. On the **Descriptive Fields** page, verify the descriptive fields information. This also should be correct because it is based on the definition you created during the conversion session. (However, note that if you defined any of the fields as picklist type, you need to create the picklist of available choices in the **Picklist** box by clicking the **New (Insert)** button and then typing the first selection in the box. Repeat this until all the picklist items have been added, and click **OK**.) When you are satisfied with your descriptive field selections, click **Next**.
9. On the **Entry Structure** page, review the entry structure of your termbase. Again, this should be correct because it is based on the definition you created during the conversion session. Here, you can also define each individual descriptive field as **Mandatory** (the field has to appear at that level at least once in every entry) or **Multiple** (the selected field can appear several times at that level in every entry) under the **Field settings**, if needed. When finished, click **Next**. When the **Wizard Complete** page comes up, click **Finish**.

You now have a new empty termbase that is open in MultiTerm. Next you need to import terms (your converted data) into the termbase. *See Step 3: Import terms.*

Step 3: Import terms

1. In MultiTerm, select **Termbase > Import Entries**. This opens the **Import** tab in the **Termbase Catalogue** page. Click **Process** (not **OK**).
2. Click **Browse** to select the XML file (= termbase data file) you created in the conversion session (Step 1). The **Log file** information is automatically filled out.
3. Select the **Fast import** option and click **Next**.
4. Click **Next** on the **Import Definition Summary** page and check how many "entries were processed". Click **Next** > **Finish**. That will take you back to the **Import** tab of the **Termbase Catalogue** dialog box. Click **OK**.

That's it! The first imported entry should now be displayed in the entry pane of MultiTerm (see Figure 4). If you want to use your MultiTerm termbase with Trados, see *Using Trados Term Recognition Feature* in the previous MultiTerm article (January 2009).

Exchanging Termbases with Others

There are two different ways to exchange termbases: either by creating and loading an MDB file, or by using XDT (termbase definition) and XML (termbase data) files. If you are using MultiTerm 7.x it's easiest to exchange termbases as MDB files, as follows:

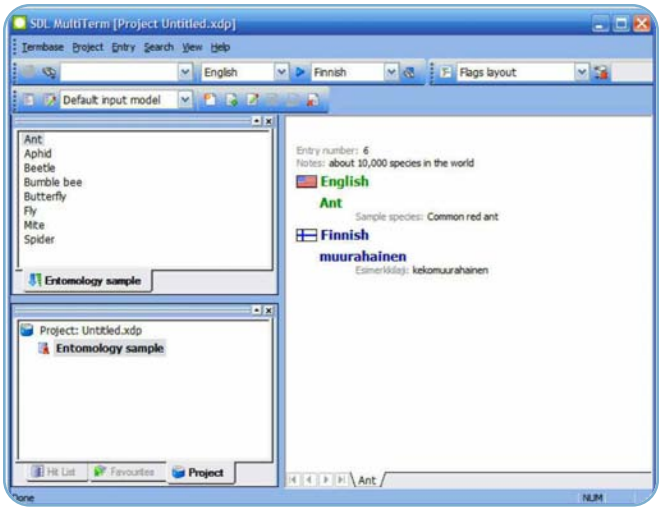


Figure 4. Our converted sample glossary as it appears in MultiTerm with the first term "Ant" displayed in the Entry pane and the other entries listed in the smaller Browse pane on the left. Note the location of the three Descriptive fields in the open entry.

Giving a termbase (.mdb file) to someone else*

Create an MDB file:

- Select **Termbase > Package/Delete Termbase**.
- Select the desired termbase from the list and name the new MDB file by selecting the **Package the termbase to this file** option, clicking **Browse** and then selecting the target folder and entering the name for the MDB file in the **File name** box. Click **Save**. Make sure that the **Delete termbase permanently** option is not selected unless you really want to delete the termbase. Click **OK**. Answer **Yes** to the annoying "**Are you sure you want to do this?**" question that pops up.
- Note that MultiTerm does not give any confirmation or indication that the process has succeeded. The only way to find out is to check that the new MDB file was created in the folder you specified. Give this MDB file to the person with whom you want to share the termbase.

Receiving a termbase (.mdb file) from someone else*

If you receive an MDB file you need to load it in order to have the termbase available to you. Load the file as follows:

- Select **Termbase > Load External Termbase**.
- Locate the desired termbase (MDB file) by clicking **Browse**, select the file and click **Open**. The file name and path appear in the **Termbase location** box.
- Name the new termbase in the **Termbase name** text box, and add a **Termbase description**, if needed. Click **OK**.
- Next you are offered an option to delete the MDB file after it has been loaded. Answer **Yes** or **No** depending on whether you want to save it.
- You do not get any other indication about the process, but the new termbase should be now available in your termbase list, which you can access normally by selecting **Termbase > Open/Close Termbases**.

So, now you should be able to create or load a termbase and use it with the automatic Term Recognition feature while translating with Trados. In my next article, I will explain how to enter new terms directly from Word and TagEditor during translation, and some other features that will make your MultiTerm experience even more beautiful. ▢

Tuomas Kostiainen (tuomas@jps.net) is an English to Finnish translator and Trados trainer, and has given several Trados workshops and presentations. For more Trados help information, see www.finntranslations.com/tradoshelp.

Register for the Mailing List

If you haven't already done so, be sure to subscribe to the LTD mailing list. Go to the Division's website (http://www.ata-divisions.org/LTD/) and click on "LTD Mailinglist." Our listmaster, Katrin Rippel, can't wait to hear from you!

Product Survey

Naomi Sutcliffe de Moraes

A Survey of Corpus Tools for Translators —in the Words of the Vendors Themselves!

Let me begin by defining some terms. When most people think of translation tools, they think of a translation environment tool using a Translation Memory (TM).

- A **translation environment tool** is a tool which "imports" your source file in some way, then leads you sentence by sentence through it, providing a field or cell in which to type the translation, then "exports" the target file in some way so that the layout of the finished translation mimics that of the original. Each tool performs these steps differently.

Corpus tools for translators allow you to search a parallel aligned bilingual corpus and a terminology database using the same interface.

- **Translation Memory**, commonly called **TM**, is a database of sentences from prior translations, linked with their translations. Tools working with TMs can store them however they wish (often in proprietary formats, or in the standard TMX format). These databases usually do not contain much context information—sometimes a code indicating the client, the field, or the translator. The sentences are all mixed up in the database, so later the translations may make no sense out of context.

Another, less-known tool for translators is a **terminology database**. Most translation environment tools have one built in, or one that is separate but compatible. They allow you to input at least the source and target terms, and sometimes much more information, such as client, field, synonyms, definition, even images.

There is, however, a third kind of tool that incorporates facets of all three of the above types of translator tools, the **corpus tool**. What is a **corpus**[1], you may ask?

A corpus is a collection of texts in electronic format. They come in many flavors:

- **Monolingual corpus** (aka **reference texts**) —many texts in a single language
- **Bilingual corpus**—many texts in two languages
  - Parallel aligned bilingual corpus (aka bitexts)—source texts and their translations, aligned for comparison purposes. The information stored is similar to that stored in a TM, but the files are stored as a whole; so when looking up a word or a sentence, you have access to the entire document as context.

The **corpus tools** for translators described below allow you to search a **parallel aligned bilingual corpus**—which they call by different names—and a terminology database using the same interface. You can search all files or just a subset of them. They also all provide automatic alignment tools which are preferable to the manual alignment required by most **translation environment tools**, which assume you will populate the TM while translating in the tool's environment, rather than by importing translations done outside it. They are extremely useful where **translation environment tools** usually fall short—when the text to be translated is in a format that cannot be imported. Examples are paper documents and scanned pdfs.

I asked four corpus tool vendors to answer the following questions:

- How does working with a corpus-based tool differ from working with a TM-based tool?
- What are the advantages?
- How does your tool use corpora?

Their answers are printed below. The main features of each tool are:

FIND by Beetext:

- Find is not an **environment tool**, but it can work as part of a software suite that includes a **translation environment tool** called Echo.
- Find searches for terms in a **terminology database** and in your **bitexts** from the same interface, displaying all results on one page.

1 Note that the plural of corpus is corpora.

LogiTerm by Terminotix:

- LogiTerm is not an **environment tool**, but it does perform preprocessing on text files, inserting bitext and terminology matches into a copy of the source file. It calls this feature LogiTrans, but it is part of LogiTerm.
- LogiTerm searches for terms in a **terminology database**, in your **bitexts** and in **reference texts** from the same interface. LogiTermWeb displays all results on one page, while the desktop version shows results in three different pages.

MultiTrans by MultiCorpora:

- MultiTrans is both a **translation environment tool** that searches bitexts to find matches for segments to be translated and a stand-alone **corpus** and **terminology database** search tool.
- MultiTrans, as a **translation environment tool**, provides matching below the segment level, at subsegment (phrase) level.
- MultiTrans searches for terms in a **terminology database** and in your **bitexts** from the same interface, displaying all results on different tabs.
- MultiTrans' **parallel corpus** may include more than two languages (a tritext, quadritext, etc.)

Transit NXT by STAR:

- Transit NXT is both a **translation environment tool** that searches bitexts to find matches for segments to be translated and a stand-alone **corpus** and **terminology database** search tool.
- Transit NXT has a function in addition to the standard source-language concordance search that searches **bitexts** (both source and target languages) for translated pairs of words or phrases.
- Transit NXT automatically searches **bitexts** (both source and target languages) for matches.
- Transit NXT's **parallel corpus** may include more than two languages (a tritext, quadritext, etc.)

Read the following descriptions, provided by the vendors themselves, and see what these corpus tools offer. Where the vendors' terms differ from those used above, I have added my standard terminology as a guide.

Beetext FIND

FIND Desktop is a search engine for translation professionals to look up terms in their own translation archives [bitexts], featuring a visual bitext display function and a user-friendly terminology management interface. The bitext function allows users to browse the source and target documents side by side and re-use previously translated phrases, sentences or entire paragraphs in any translation environment, such as a TM or a text editor. The lexicon allows you to create terminological entries directly from the bitext display or manually. Lexicons can be exported in a .CSV format, from which they can be imported into a spreadsheet program such as Excel, into a translation memory or for sharing terminology with a colleague that uses FIND Desktop as well. Beetext FIND is an affordable, low-maintenance tool that can be used with any document type, repetitive or not. It only takes a few minutes to get started and start getting payoff from your investment.

Document Formats

FIND recognizes most document formats, such as WordPerfect, Word, PDF, PowerPoint, Excel, html and more than 200 other formats. In the case of PDF files, it should be noted that scanned documents are an image of a text and, thus, FIND will not be able to extract text from them.

Automatic Bitext Display

When a search is initiated, results from the bitext search, as well as the lexicon, are automatically displayed. Each section contains its own result list. Search results are returned in a flash, and the searched terms will be highlighted in the text shown by FIND.

Figure 1 (next page) is a screen capture of a result matched to several documents. Each component is described below.

- The **Previous** and **Next** buttons at the bottom left of the screen will toggle from one occurrence to another in the selected document. The corresponding document will scroll accordingly.
- The **Previous** and **Next** buttons at the bottom center will toggle from one document to another in the result list.
- The **Previous** and **Next** buttons at the bottom right of the screen will toggle from one corresponding version of the document to another. For example, if an English document is matched to a Spanish version, a French

version and a German version, the right side buttons become active and allow you to switch from one version to another.

- The document names are displayed as hyperlinks under the texts. This link will open the original document when clicked.
- The **Create Entry** button allows you to automatically create a new entry in the lexicon from the bitext. Just highlight an expression in both texts and click on the **Create Entry** button. The highlighted expression in the left side window will appear in the Term field of the new entry, and the phrase on the right side will appear in the Equivalent field. FIND will also automatically extract a context and place it in the **Context** field for each version.

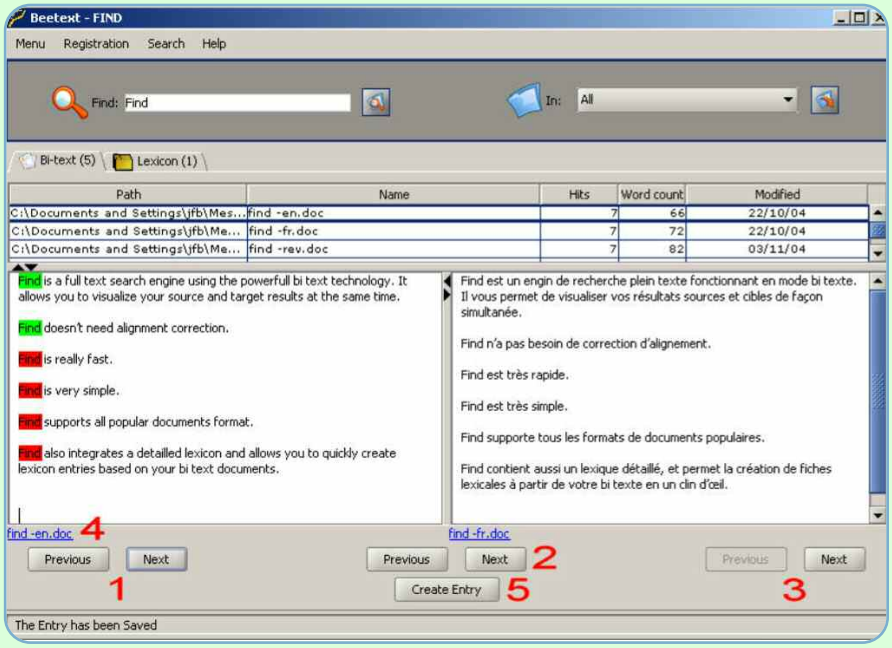

Beetext - FIND
Menu Registration Search Help
Find: Find
In: All
Bi-text (5) Lexicon (1)
Path Name Hits Word count Modified
C:\Documents and Settings\jfb\Mes... find -en.doc 7 66 22/10/04
C:\Documents and Settings\jfb\Me... find -fr.doc 7 72 22/10/04
C:\Documents and Settings\jfb\Me... find -rev.doc 7 82 03/11/04
Find is a full text search engine using the powerfull bi text technology. It allows you to visualize your source and target results at the same time.
Find doesn't need alignment correction.
Find is really fast.
Find is very simple.
Find supports all popular documents format.
Find also integrates a detailled lexicon and allows you to quickly create lexicon entries based on your bi text documents.
Find est un engin de recherche plein texte fonctionnant en mode bi texte. Il vous permet de visualiser vos résultats sources et cibles de façon simultanée.
Find n'a pas besoin de correction d'alignement.
Find est très rapide.
Find est très simple.
Find supporte tous les formats de documents populaires.
Find contient aussi un lexique détaillé, et permet la création de fiches lexicales à partir de votre bi texte en un clin d'oeil.
find -en.doc
find -fr.doc
Previous Next
Create Entry
The Entry has been Saved
1
2
3
4
5

Figure 1: A result matched to several documents in Beetext FIND.

Integrated Lexicon [*Terminology Database*]

FIND's lexicon is user-friendly and fast. Here is an overview of the lexicon interface and a description of its components. (Figure 2)

The following features are shown in Lexicon mode:

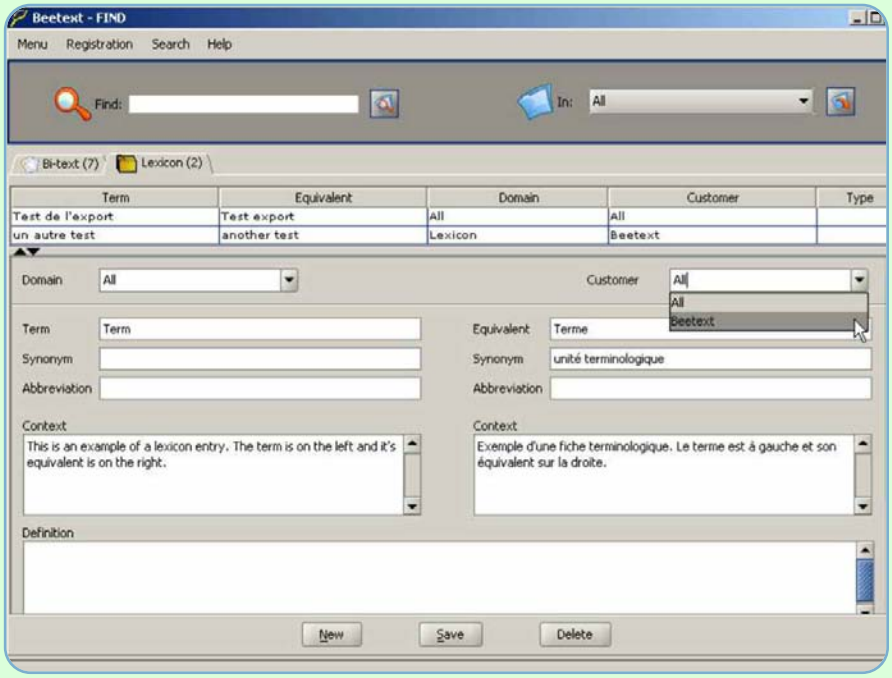


Figure 2: Lexocon Interface, Beetext FIND.

- **Result List**: The lexicon result list works like the bitext result list. The last column Type contains the field in which the term or expression was found.
- The **Domain** and **Client Fields**: In these fields you can select the domain and the client related to the entry. You can either choose from the drop-down menu or type a new entry in the field. Once you have entered a new entry, it is added to the list for future reference.
- **Synonym** and **Abbreviation Fields**: These fields contain a synonym and an abbreviation for the term and for its equivalent. These fields are included in the search.
- **Context Fields**: These fields contain a context for the term and its equivalent, which are displayed in bold in the context. The context field is filled automatically when an entry is created from a bitext and the original file name is displayed under the context field.
- **Notes** and **Definition Fields**: Enter the definition and personal notes in these fields.

For more information about FIND Desktop or Server edition, please visit www.beetext.com.

Terminotix Inc.

The term **translation memory** originally referred to a tool that stored source and target segments in a database—a black box with little flexibility. Nowadays, translation memory tools are more flexible, especially in terms of displaying the context of a segment rapidly when pretranslating or searching; however, their architecture is still not as flexible as a product like LogiTerm and its automatic retrieval tool, LogiTrans. Let me explain why.

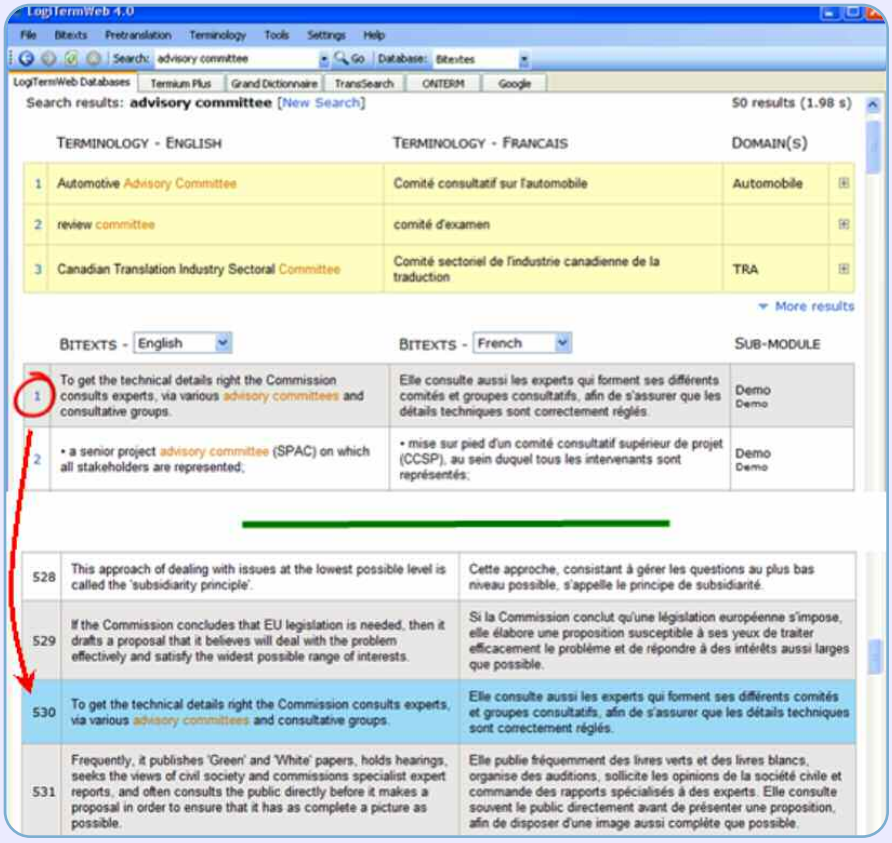

LogiTermWeb 4.0
File Bitexts Pretranslation Terminology Tools Settings Help
Search: advisory committee Go Database: Bitexts
LogiTermWeb Databases Termium Plus Grand Dictionnaire TransSearch ONTERM Google
Search results: advisory committee [New Search]
50 results (1.98 s)
TERMINOLOGY - ENGLISH
TERMINOLOGY - FRANCAIS
DOMAIN(S)
1 Automotive Advisory Committee
Comité consultatif sur l'automobile
Automobile
2 review committee
comité d'examen
3 Canadian Translation Industry Sectoral Committee
Comité sectoriel de l'industrie canadienne de la traduction
TRA
More results
BITEXTS - English
BITEXTS - French
SUB-MODULE
1 To get the technical details right the Commission consults experts, via various advisory committees and consultative groups.
Elle consulte aussi les experts qui forment ses différents comités et groupes consultatifs, afin de s'assurer que les détails techniques sont correctement réglés.
Demo
2 • a senior project advisory committee (SPAC) on which all stakeholders are represented;
• mise sur pied d'un comité consultatif supérieur de projet (CCSP), au sein duquel tous les intervenants sont représentés;
Demo
528 This approach of dealing with issues at the lowest possible level is called the 'subsidiarity principle'.
Cette approche, consistant à gérer les questions au plus bas niveau possible, s'appelle le principe de subsidiarité.
529 If the Commission concludes that EU legislation is needed, then it drafts a proposal that it believes will deal with the problem effectively and satisfy the widest possible range of interests.
Si la Commission conclut qu'une législation européenne s'impose, elle élabore une proposition susceptible à ses yeux de traiter efficacement le problème et de répondre à des intérêts aussi larges que possible.
530 To get the technical details right the Commission consults experts, via various advisory committees and consultative groups.
Elle consulte aussi les experts qui forment ses différents comités et groupes consultatifs, afin de s'assurer que les détails techniques sont correctement réglés.
531 Frequently, it publishes 'Green' and 'White' papers, holds hearings, seeks the views of civil society and commissions specialist expert reports, and often consults the public directly before it makes a proposal in order to ensure that it has as complete a picture as possible.
Elle publie fréquemment des livres verts et des livres blancs, organise des auditions, sollicite les opinions de la société civile et commande des rapports spécialisés à des experts. Elle consulte souvent le public directement avant de présenter une proposition, afin de disposer d'une image aussi complète que possible.

Figure 3: Bitext search results shown in context, Logiterm.

Document-based [*Corpus-based*] vs. Segment-based

The document concept is central to LogiTerm. Aligning a pair of documents produces a bitext file, which is a self-contained HTML file that displays source and target texts side by side, with segments aligned in the order in which they originally appeared (linear alignment). A bitext file can be saved to any disk location; anyone can view it and perform searches without using LogiTerm.

WYSIWYG File Management

One advantage of using LogiTerm is that a translator who stores client files in Windows folders can continue working in exactly the same way, because bitext files are stored by default in the same folder as the original files. The following example shows files containing a source text, bitext and target text:

Modules are bitext groupings created in LogiTerm that act similarly to a translation memory. A module is actually a "proxy" pointing to folders that contain bitext files.

Bitext files stay in their original locations and are always available. The contents of a module can be determined by browsing its folders. To change the contents of any module, you simply change the files in its folders or edit the contents of the bitext files. Everything is open and visible. There is no black box, as is the case with most translation memories.

Context is Just a Click Away

Bitexts are complete documents, so when you search a LogiTerm module, the context of each result segment is just a click away. If you click on bitext result "1" (Figure 3), that result is shown in the context of the original file from which it was taken.

This is possible because segments are aligned linearly and bitexts are always kept in individual files (they are never combined as with translation memories).

"Best Match" Source Text Analysis

Traditional translation memories are like collections of sentences, while LogiTerm modules are more like collections of documents. The analysis performed by LogiTrans automatically gives priority to the bitexts that are most similar overall to the source text, and shows these results first, because the translations in those bitexts will likely be more relevant. Additionally, the translations retrieved will be more consistent, since they come from a smaller number of bitexts.

Furthermore, you can set up LogiTrans to use a specific bitext as a data source, just as you would a LogiTerm module. No preparation is required. Translations in selected bitexts are given priority over

translations in any selected modules.

Using Monolingual Reference Documents [*reference texts*]

Another advantage of using LogiTerm is that LogiTrans can analyze the similarity of unilingual reference documents. You can identify similarities to documents for which bitexts have not yet been created, but for which translations are available.

Finally, monolingual reference documents from clients or other sources can be searched manually in LogiTerm, and thereby provide valuable information.

Data-friendly Approach

Aside from the "best match" capability mentioned above, most of the strengths of LogiTerm and LogiTrans can be summed up in one expression: "data friendliness." LogiTerm allows you to search in documents of many different types. Data friendliness means less preprocessing or conversion work for you, and thus you can get the most out of your data in a wide range of situations.

For more information about LogiTerm, visit www.terminotix.com.

MultiTrans

MultiCorpora

Working with a corpus-based tool differs significantly from working with TM-based products. In fact, our technology alone is so different from classic TM systems that the Translation Automation User Society (TAUS) has coined our Technology "Advanced Leveraging TM."

Some of the inherent advantages:

- Alignment benefits
  - Provides context for each segment at paragraph level
  - Produces quality 1:N and N:1 alignments rapidly
  - Creates multi-directional, multilingual TMs [**bitexts, or parallel aligned bilingual corpora**]
  - Delivers ability to create a fully useable 10,000 segment TM [**parallel corpus**] in under five minutes
- Advanced Leveraging Functionality
  - Identifies and replaces matching paragraphs
  - Identifies sub-segments and their translations in context (within strings and sentences)
  - Provides more repetitions, out of existing TMs [bitexts]
- Preserve the full context of every segment, even at the entire document level
- Interactive translation module allows direct view of context

A classical translation memory separates your document into out-of-context sentences and then tries to align them. In layperson's terms, it can be compared to an Excel worksheet, where, for instance, in one column you have the English, and in another column the French equivalent sentences. It is quite tedious to verify that each sentence or sentence group is perfectly aligned and because the documents are split into individual sentences, the context is lost.

The MultiTrans TextBase TM [parallel corpora] indexes your integral documents in a side by side manner. Since the technology does not divide documents into individual sentences, MultiTrans can easily align paragraphs and sentences at a near perfect level without any human intervention. Furthermore, the context of your past translations (the entire document) is preserved. This means that in the rare cases of misalign-

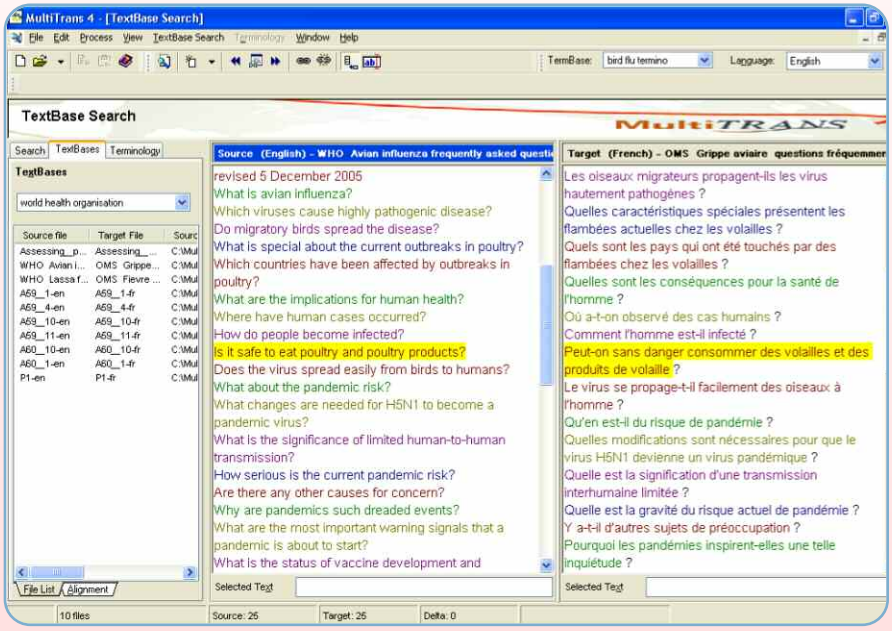

MultiTrans 4 - [TextBase Search]
File Edit Process View TextBase Search Terminology Window Help
TermBase: bird flu termino
Language: English
TextBase Search
MultiTRANS
Search TextBases Terminology
TextBases
world health organisation
Source file
Target File
Source (English) - WHO Avian influenza frequently asked questi
Target (French) - OMS Grippe aviaire questions fréquemmer
revised 5 December 2005
What is avian influenza?
Which viruses cause highly pathogenic disease?
Do migratory birds spread the disease?
What is special about the current outbreaks in poultry?
Which countries have been affected by outbreaks in poultry?
What are the implications for human health?
Where have human cases occurred?
How do people become infected?
Is it safe to eat poultry and poultry products?
Does the virus spread easily from birds to humans?
What about the pandemic risk?
What changes are needed for H5N1 to become a pandemic virus?
What is the significance of limited human-to-human transmission?
How serious is the current pandemic risk?
Are there any other causes for concern?
Why are pandemics such dreaded events?
What are the most important warning signals that a pandemic is about to start?
What is the status of vaccine development and
Les oiseaux migrateurs propagent-ils les virus hautement pathogènes ?
Quelles caractéristiques spéciales présentent les flambées actuelles chez les volailles ?
Quels sont les pays qui ont été touchés par des flambées chez les volailles ?
Quelles sont les conséquences pour la santé de l'homme ?
Où a-t-on observé des cas humains ?
Comment l'homme est-il infecté ?
Peut-on sans danger consommer des volailles et des produits de volaille ?
Le virus se propage-t-il facilement des oiseaux à l'homme ?
Qu'en est-il du risque de pandémie ?
Quelles modifications sont nécessaires pour que le virus H5N1 devienne un virus pandémique ?
Quelle est la signification d'une transmission interhumaine limitée ?
Quelle est la gravité du risque actuel de pandémie ?
Y a-t-il d'autres sujets de préoccupation ?
Pourquoi les pandémies inspirent-elles une telle inquiétude ?
Selected Text
Selected Text
File List Alignment
10 files
Source: 26
Target: 26
Delta: 0

Figure 4: TextBase TM: Full context at the entire document level

ment, you can easily realign a sentence on-the-fly, while you are translating! Note that the MultiTrans TextBase TM can be multilingual and multidirectional, and therefore eliminates the need for exponentially duplicating bilingual memories.

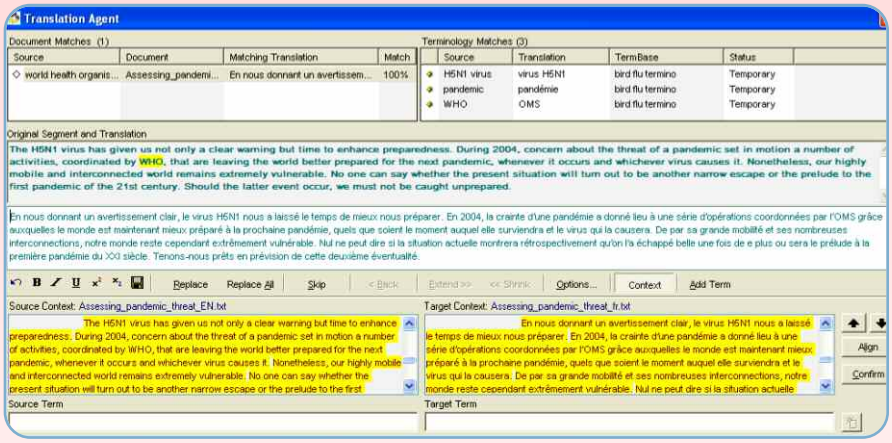

Translation Agent
Document Matches (1)
Source
Document
Matching Translation
Match
world health organis...
Assessing_pandemi...
En nous donnant un avertissem...
100%
Terminology Matches (3)
Source
Translation
TermBase
Status
H5N1 virus
virus H5N1
bird flu termino
Temporary
pandemic
pandémie
bird flu termino
Temporary
WHO
OMS
bird flu termino
Temporary
Original Segment and Translation
The H5N1 virus has given us not only a clear warning but time to enhance preparedness. During 2004, concern about the threat of a pandemic set in motion a number of activities, coordinated by WHO, that are leaving the world better prepared for the next pandemic, whenever it occurs and whichever virus causes it. Nonetheless, our highly mobile and interconnected world remains extremely vulnerable. No one can say whether the present situation will turn out to be another narrow escape or the prelude to the first pandemic of the 21st century. Should the latter event occur, we must not be caught unprepared.
En nous donnant un avertissement clair, le virus H5N1 nous a laissé le temps de mieux nous préparer. En 2004, la crainte d'une pandémie a donné lieu à une série d'opérations coordonnées par l'OMS grâce auxquelles le monde est maintenant mieux préparé à la prochaine pandémie, quels que soient le moment auquel elle surviendra et le virus qui la causera. De par sa grande mobilité et ses nombreuses interconnections, notre monde reste cependant extrêmement vulnérable. Nul ne peut dire si la situation actuelle montrera rétrospectivement qu'on l'a échappé belle une fois de plus ou sera le prélude à la première pandémie du XXI siècle. Tenons-nous prêts en prévision de cette deuxième éventualité.
Replace
Replace All
Skip
< Back
Extend >>
<< Shrink
Options...
Context
Add Term
Source Context: Assessing_pandemic_threat_EN.txt
Target Context: Assessing_pandemic_threat_fr.txt
Align
Confirm
Source Term
Target Term

Figure 5: Translation Agent: Full paragraph matches, retaining the full document context.

This philosophical difference means that, on top of being able to preserve the context, the TextBase TM technology enables you to rapidly build massive memories of legacy translations. There is no need for costly manual verification of alignments before they can be used productively.

As a result, you can create a much larger translation memory [parallel corpus] in a much shorter timeframe. On most systems, the TextBase TM can be created at the astonishing rate of 6 to 10 million words per hour! This means that instead of being limited to the size of a conventional TM, which often remains small because of the effort it takes to build, you can now index all of your legacy documentation. Having a larger pool of reliable, previously translated data to compare against means that you will find a lot more repetition. This increases quality, terminology cohesion, and productivity.

Since it is based on full texts, the TextBase TM approach also enables enhanced data mining when comparing a document against the TM. Since the TM is not segmented, mining can take place on full paragraphs. Instead of assembling a paragraph to be translated from disparate sentences that do not flow together, the TextBase TM will identify exact paragraph matches, and then return the full paragraph as a single retrieved translation segment.

Like classic TM systems, MultiTrans also identifies and replaces full and fuzzy sentences. It is, however, designed to go beyond the segment, to proactively identify sub-segments. This means that segments that fall below the fuzzy level, which are ignored by most conventional TM systems, are actually identified by MultiTrans and this increases your multilingual asset pool tremendously! In other words, with MultiTrans, you can get a lot more repetitions, more cohesiveness and greater productivity gains.

For more information about MultiTrans, visit www.multicorpora.ca.

STAR Transit NXT: It's All About the Context

While the whole idea behind corpora and translation memory (TM) is to increase translator consistency and productivity, a basic "quality" principle states that the quality of an output is relative to the quality of the input. Therefore, in translations, the quality of proposals based on TM is directly related to the quality of the reference material the TM is built upon.

In order to leverage existing translations, multilingual reference material can be created through alignment of source and target documents and files. However, when building translation reference materials, *contextually correct subject-matter alignment is critical*.

Although Transit NXT is generally classified as a translation memory tool, it embodies many of the characteristics of a corpus tool. While the classic alignment and TM tools focus on storing isolated segments, the highest quality TM systems, such as Transit NXT, will prioritize contextually relevant suggestions, because without context, errors will occur.

There are two approaches to aligning and storing source and target segments. One is to store them with context as Transit NXT does—the other is to store them as isolated segments. The dominant industry alignment practice aligns segments in isolation in a database, which leads to avoidable errors, whereas aligned segments stored with context in a file system can still refer to the original document and its context.

The following example shows how segments in isolation are stored and corresponding translations are proposed. Take note of the text highlighted in blue:

Now, if the new source text below is translated using

English (Source)	French (Target)
Please, check the existing grounding.	Veillez à vérifier la mise à la terre.
If it fails the test, please refer to a professional.	Si elle est déficiente, contacter un spécialiste.

reference segments stored in isolation, it would be translated as follows:

However, in the new context of "ground cable", the

New Source

"Please, check the ground cable. *If it fails the test, please refer to a professional.*"

Proposed Translation

"Veillez à vérifier la mise à la terre. Si elle est déficiente, contacter un spécialiste."

translation of the second sentence should actually be, "S'il est déficient, contacter un spécialiste."

Using Transit NXT, the correct translation would have been proposed because Transit NXT stores its TM with context and offers prioritized, contextually relevant translation proposals.

Figure 6 shows Transit NXT's capability to align and retrieve reference material in context. This capability is only possible because segments are stored in context as a multilingual file based TM, as opposed to being stored as isolated segments in a database:

While the above example shows typical pre-translation capability in Transit NXT, the following context sensitive functions are also available in Transit NXT in order to accelerate and improve the translator's productivity:

- **Terminology Management**: TermStar NXT, Transit NXT's terminology component, assures the correct term is always used
- **Concordance Search**: quickly searches the current project files as well as reference material for individual words, phrases or similar terms to see the context in which they are used
- **Dynamic Linking**: similar to concordance search, Dynamic Linking searches both a source and target language for translated pairs of individual words, phrases or similar terms to see the context in which they are used
- **Dual Fuzzy**: if no matches are found in the source text, Transit NXT searches the target text for similar sentences while the translation is being entered. Transit NXT also searches both the source and the target language reference materials for suggestions
- **Sync View**: provides additional context by displaying the document and software layout, as well as corresponding graphics in the Sync View window.

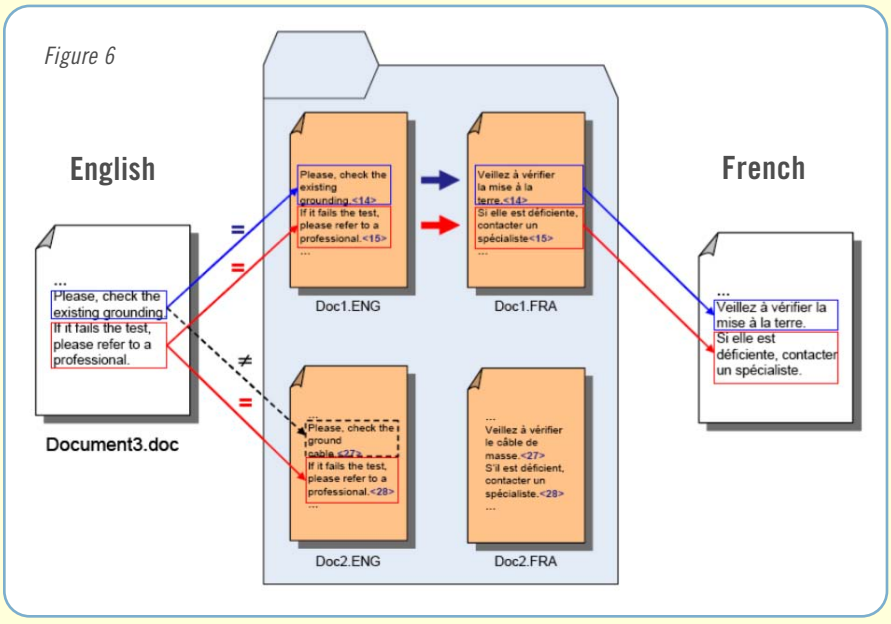


Figure 6

STAR is an industry leader that for 25 years has focused on providing the most efficient multilingual information services and solutions. STAR has developed the complete suite of technologies to fulfill that mission and makes all of its technology available to the open marketplace (Figure 7, next page).

For more information please visit: www.star-group.net. ▢

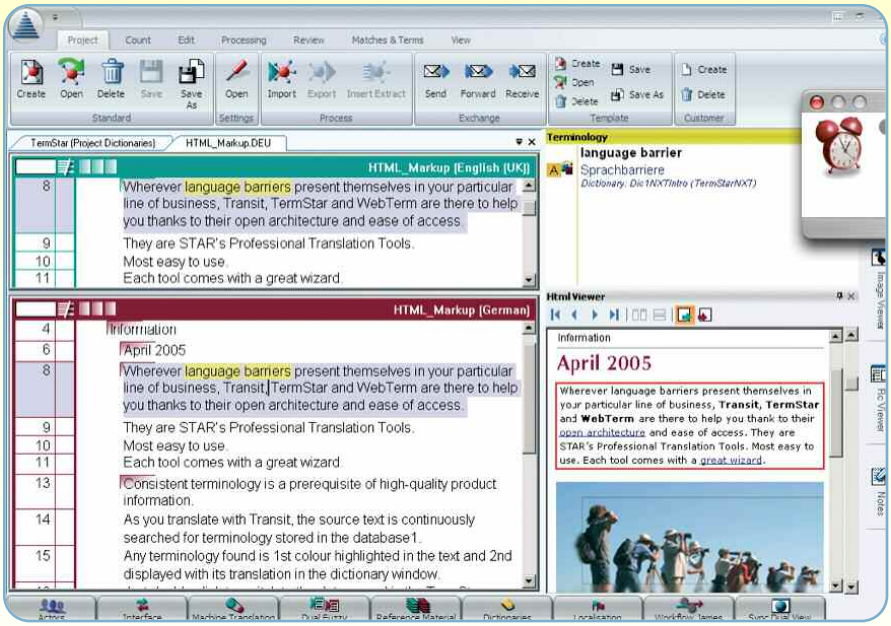

Project
Count
Edit
Processing
Review
Matches & Terms
View
Create
Open
Delete
Save
Save As
Import
Export
Insert Extract
Send
Forward
Receive
Standard
Settings
Process
Exchange
Template
Customer
TermStar (Project Dictionaries)
HTML_Markup.DEU
HTML_Markup [English (UK)]
Wherever language barriers present themselves in your particular line of business, Transit, TermStar and WebTerm are there to help you thanks to their open architecture and ease of access.
They are STAR's Professional Translation Tools.
Most easy to use.
Each tool comes with a great wizard.
HTML_Markup [German]
Information
April 2005
Wherever language barriers present themselves in your particular line of business, Transit, TermStar and WebTerm are there to help you thanks to their open architecture and ease of access.
They are STAR's Professional Translation Tools.
Most easy to use.
Each tool comes with a great wizard.
Consistent terminology is a prerequisite of high-quality product information.
As you translate with Transit, the source text is continuously searched for terminology stored in the database1.
Any terminology found is 1st colour highlighted in the text and 2nd displayed with its translation in the dictionary window.
Terminology
language barrier
Sprachbarriere
Dictionary: Dic1NXTIntro (TermStarNXT)
HtmlViewer
Information
April 2005
Wherever language barriers present themselves in your particular line of business, Transit, TermStar and WebTerm are there to help you thanks to their open architecture and ease of access. They are STAR's Professional Translation Tools. Most easy to use. Each tool comes with a great wizard.

Figure 7: Transit NXT's User Interface showing the translator's work environment, including dynamic terminology control and end layout preview (Sync View)

Further Reading:

For further reading on the subject of corpus tools, see the following two articles:

The Translator's Binoculars, *ATA Chronicle*, Part 1; August 2008, Part 2; September 2008.

For a full review of LogiTerm, see the following two articles:

LogiTerm, Your Personal Search Engine. *ATA Chronicle*, Part 1: November 2007; Part 2: January 2008.

Found CAT, cont. from page 8

As a user and volunteer localizer of OmegaT, I had a front row seat during the group's successful inclusion of the OpenOffice.org spellchecker and TMX format compatibility enhancement efforts—not to mention improved support for right-to-left and non-Latin alphabet languages.

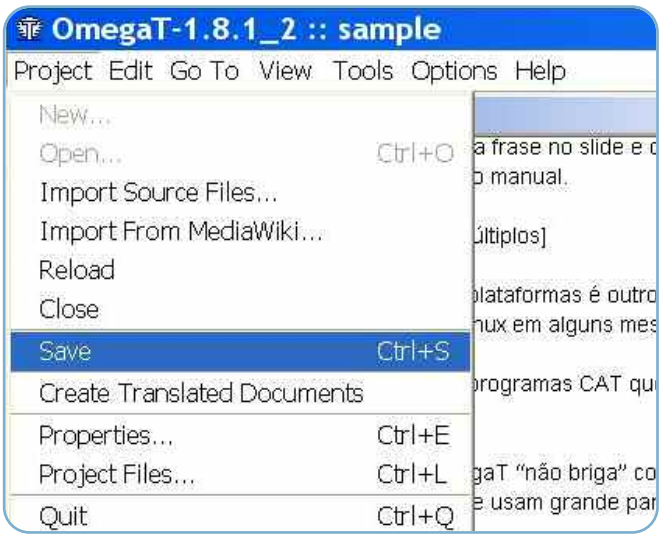


Figure 4

As I write, one of the developers just finished a script to allow OmegaT to accept Trados-generated ttx files. To convert other formats not directly supported, OmegaT also uses OpenOffice.org, Okapi Framework (Windows-only) and the Translate ToolKit.

If you are unsure about which CAT program to use, I would recommend that you try OmegaT. It is free, no strings attached. Then, if it seems like a poor fit, you can ask questions and discuss issues on the OmegaT list on Yahoo. OmegaT product improvement suggestions and new function requests should be sent to the OmegaT project at sourceforge.net.

To read more about OmegaT, please visit http://en.wikipedia.org/wiki/OmegaT or http://www.omegat.org/en/omegat.html. ▢

Call for Reviewers

Are you using software that helps you in your day-to-day work as a translator and/or interpreter? Tell us about it! Send reviews of your new, favorite, even most-hated language technology software to Roomy Naqvy at roomynaqvy@gmail.com. Your colleagues will thank you!